2015 IEEE International Conference on Multimedia Big Data

Service Objective Evaluation via Exploring Social Users' Rating Behaviors

Guoshuai Zhao
SMILES LAB
Xi'an Jiaotong University
Xi'an China
zgs2012@stu.xjtu.edu.cn

Xueming Qian*
SMILES LAB
Xi'an Jiaotong University
Xi'an China
qianxm@mail.xjtu.edu.cn

***Abstract*—With the boom of e-commerce, it is a very popular trend for people to share their consumption experience and rate the items on a review site. The information they shared is valuable for new users to judge whether the items have high-quality services. Nowadays, many researchers focus on personalized recommendation and rating prediction. They miss the significance of service objective evaluation. Service objective evaluation is usually represented by star level, which is given by a large number of users. The more user ratings, the more objective evaluation is. But how does it work for new items? It is lack of objectivity if there are few users have rated to the item, such as there are just two ratings. In this paper, we propose a model to solve service objective evaluation by deep understanding social users. As we know, users' tastes and habits are drifting over time. Thus, we focus on exploring user ratings confidence, which denotes the trustworthiness of user ratings in service objective evaluation. We utilize entropy to calculate user ratings confidence. In contrast, we mine the spatial and temporal features of user ratings to constrain confidence. We conduct a series of experiments based on Yelp datasets. Experimental results show the effectiveness of proposed model.**

Keywords- recommender system; service objective evaluation; user ratings confidence; social networks

I. Introduction

Recently people receive more and more digitized information from Internet. The volume of information is larger than any other point in time, reaching a point of information overload. A great deal of information fills the Internet, so that we are confused about authenticity and objectivity of information. Especially when we choose an item, we will heavily rely on the already accumulating comments and ratings. But for new items, there are just few comments and ratings. Additionally, the objectivity of comments and ratings is not guaranteed. Suppose that, there are two restaurants, and one of the two restaurants is strictly better. However, for some reason, the first customers give lower ratings to this high quality restaurant. Then other customers, who rely on ratings shown in website to make their choice, will make the wrong decision. From another aspect, it will make customers confused if there are just two contrary ratings to an item. For example, there exists a new item (the right service shown in Fig. 1), named *Cafe*, and it just has two already accumulating comments and ratings. One user rated it 2 stars, and another rated it 5 stars. Then which one we should trust? Or whose rating is more confidence? Factually, official website generally computes the average ratings, and sets it as star level to each item. It is an apposite approach for items which have been rated by large number of users. But for a new item, we cannot straightforwardly see the few ratings as the objective evaluation to this item.

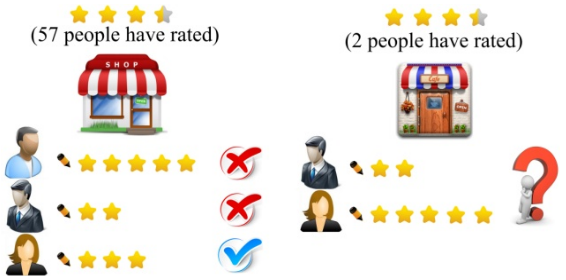


Figure 1. An Example for Motivation.

However, researches mostly focus on personalized recommendation and rating prediction. The first generation of recommender systems [10] with traditional collaborative filtering algorithms [11]-[25], and many social network based models [26]-[39] mostly aims at recommending personalized services, or predicting user preferences and ratings. They miss the significance of service objective evaluation. Thus, in this paper, we propose the issue of service objective evaluation, and try our best to solve it.

With this motivation, in this paper, we focus on user ratings confidence to discriminate ratings to conduct service objective evaluation. Shown as the left service in Fig. 1, we can learn user ratings confidence from training set. Additionally, we explore user ratings confidence with combining spatial-temporal features of ratings to deep understand social users. Through the approach we proposed, we can learn the confidence value of a rating within specific spatial-temporal context.

Specifically, we conduct service objective evaluation by deep understanding social users with exploring user ratings confidence. Firstly, we utilize information entropy to calculate user ratings confidence because entropy is the measure of the disorder or randomness of energy and matter in a system. Secondly, we mine spatial and temporal features of ratings from training set by observational learning to constrain user ratings confidence. At last we fuse them into a unified probabilistic model.

*Corresponding author

978-1-4799-8688-0/15 $31.00 © 2015 IEEE
DOI 10.1109/BigMM.2015.67

The biggest difference between our approach and related works is that: they focus on personalized rating prediction or recommendation, but in this paper, we focus on service objective evaluation. Our goal is to predict service overall and objective star level evaluation with few ratings.

The main contributions of this paper are as follows:

- We propose an issue about service overall and objective evaluation, and utilize biases and traditional rating prediction methods to solve it.
- We use information entropy value to compute user ratings' confidence. Furthermore, we mine ratings' features from spatial-temporal information, and find that the spatial-temporal features of users' ratings are helpful to constrain user ratings confidence.
- We propose a novel model to evaluate services by deep understanding social users with exploring user ratings confidence with combining ratings' spatial-temporal features. Experimental results show outstanding performance of our model.

II. Related Work

There are some traditional approaches could be utilized to solve the problem of service objective evaluation. The first primitive approach is using biases. Biases could represent users' rating habits, such as user A's ratings are almost 4 stars and 5 stars, while B's ratings are almost 3 stars. Koren [1] supposes customer preferences for products are drifting over time, and proposes collaborative filtering model with temporal dynamics. He considers user and item time changing biases, and compares the ability of various suggested baseline predictors. Dror *et al.* [2] propose a model, which incorporates a rich bias model with terms that capture information from the taxonomy of items and different temporal dynamics of music ratings. Even above authors focus on personalized rating prediction, we can refer the idea of user biases and taxonomy biases. Furthermore, we could convert personalized rating prediction to service objective evaluation.

There are some more approaches to predict users' ratings. A typical model is matrix factorization model. Many systems [3]-[9] employ matrix factorization techniques to learn the latent features of users and items, and predict the unknown ratings using these latent features. R. Salakhutdinov et al. [6] present the Probabilistic Matrix Factorization model which scales linearly with the number of observations and, more importantly, performs well on the large, sparse, and very imbalanced Netflix dataset. Yang et al. [4] propose to use the concept of 'inferred trust circle' based on the domain-obvious of circles of friends on social networks to predict users' ratings. Qian et al. [8] and [9] consider more social factors, including interpersonal influence, interpersonal interest similarity and personal interest based on matrix factorization to predict users' ratings. For service objective evaluation, we could use matrix factorization model to learn user and item latent features, and then predict all users' ratings to each item. The overall evaluation can be calculated by averaging the predicted ratings.

From another side, we can not only utilize users' ratings to conduct service objective evaluation, but also directly exploit similarity between items to predict evaluation. Sarwar et al. [12] propose an item-based collaborative filtering algorithm. They focus on producing the rating from a user to an item based on the average ratings of similar or correlated items by the same user. It is the one of the most popular algorithms in recommender system.

In this paper, we propose an issue about service overall and objective evaluation. To solve this problem, we introduce traditional rating prediction methods, including methods based on biases and based on matrix factorization model. Here, we list compared methods as follows:

- **BM (Basic Method)**: This method operates average rating to evaluate items directly.
- **Biases (Basic Biases)**: This method considers users' rating biases to overcome different rating criteria.
- **BT (Biases Based on Taxonomy)**: This method explores users' rating criteria with more refinements. It considers taxonomy information based on biases.
- **BaseMF (Basic Probabilistic Matrix Factorization)**: This model is basic matrix factorization approach proposed in [6] without consideration of any social factors. Once we get the learned user and item features, we can use them to predict all users' ratings to each item. Then we evaluate star level of each item by averaging predicted ratings.
- **CircleCon:** This approach proposed in [4] focuses on the factor of interpersonal trust in social network and infers the trust circle based on matrix factorization. We can utilize predicted ratings to evaluate items objectively and overall by averaging them.
- **PRM:** This approach proposed in [8] and [9] considers more social factors, including interpersonal influence, interpersonal interest similarity and personal interest based on matrix factorization to predict users' ratings. We can utilize predicted ratings to evaluate items objectively and overall by averaging them.
- **Item_based:** This approach proposed in [12] focuses on producing the rating from a user to an item based on the average ratings of similar or correlated items by the same user. We utilize its adjusted cosine similarity algorithm to compute similarity between items, and evaluate items objectively.

III. Datasets Introduction

Yelp is a local directory service with social networks and user reviews. It is the largest review site in America and it has more than 71 million monthly unique visitors as of January 2012. Yelp was founded in 2004, and included the restaurant, shopping center, hotel and tourism businesses, etc. Users can rate the businesses, submit comments, communicate shopping experience, etc. It combines local reviews and social networking functionality to create a local online community. Compared with the traditional review sites, Yelp has some characteristics. Firstly, real users' comments. Especially Yelp attracts some zealous users to their community. Secondly, Yelp encourages user interactions through various forms, and pays a good reward to the active users. Essentially, this form of intelligence allows people to actively participate and share their knowledge with other users, especially their friends.

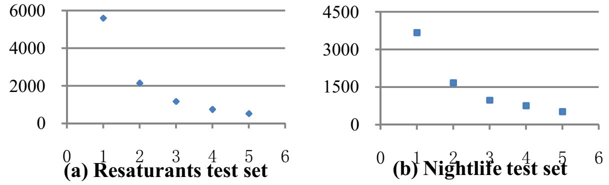
We have crawled nearly 60 thousand users' circles of friends and their rated items from November 2012 to January 2013. The disposal data consists two categories: Restaurants and Nightlife. The former dataset contains 263,124 ratings from 4,138 users who have rated a total of 62,221 items. The later dataset contains 436,301 ratings from 11,152 users who have rated a total of 21,647 items. Table 1 and Table 2 shows statistic of Yelp Restaurants dataset and Yelp Nightlife dataset respectively.

TABLE I. STATISTIC OF YELP RESTAURANTS DATASET

User number	4,138	
Item number	62,221	52,071 (training)
		10,150 (test)
Ratings number	263,124	244,205 (training)
		18,919 (test)
Average rating	3.646	

TABLE II. STATISTIC OF YELP NIGHTLIFE DATASET

User number	11,152	
Item number	21,647	14,066 (training)
		7,581 (test)
Ratings number	436,301	420,790 (training)
		15,511 (test)
Average rating	3.589	



(a) Resaturants test set

(b) Nightlife test set

Figure 2. The distributions of items in test set according to the number of ratings. Fig. (a) and (b) show the distribution of the number of items based on Yelp Restaurants and Nightlife respectively.

Note that, the issue we proposed in this paper is service objective evaluation, especially for services with few ratings. Thus, we must handle our dataset to extract appropriate test data. As shown in Table 1 and Table 2, we split items in two groups. One is training set and another is test set. The point is that each item in test set has few ratings, which are no more than 5. Fig. 2 shows the distributions of items in our two test sets according to the number of ratings. Fig. 2(a), and (b) show the distribution of the number of items based on Yelp Restaurants and Nightlife test set respectively. The y-axis represents the count of items. The x-axis represents the number of ratings under each item. For example, in Restaurants test set, there are 5605 test items just have only one rating. From Fig. 2, we can see none of these test items have more than 5 ratings.

IV. THE APPROACH

In this paper, we objectively estimate items star levels by exploring user ratings confidence. We use information entropy value to compute user ratings' confidence. Furthermore, we mine ratings' features from spatial-temporal information, which is employed to constrain user ratings confidence. The basic idea is that users' profiles are changing. That is to say user ratings confidences are different in different places at different times. At last, we combine user ratings confidence and spatial-temporal features to integrate into a unified probabilistic model. Hereinafter we turn to details of our methods.

A. User Ratings Confidence

As mentioned before, we focus on user ratings confidence to discriminate their ratings to conduct service objective evaluation. Our basic idea is that user ratings have different confidence. Then how should we know which people is trustworthy? We have large records of users' historical ratings. As shown in Fig. 1, we can exploit these large data to judge user ratings confidence. As we all know, entropy is the measure of the disorder or randomness of energy and matter in a system. If a user's ratings are confidence, his/her ratings must have little differences with items' real star levels. Thus the information entropy value of these differences can be used to represent the confidence value of user ratings. That is to say, we set the differences between user ratings and items' real star levels as an error value system, then entropy of this system can reflect user's rating habits and stability. Additionally, we add a coefficient to distinguish weights of different error values to enhance user ratings confidence, because entropy algorithm cannot make a difference in different error values. We know that, the lower entropy value is, the more stability system is. So is user ratings confidence. We represent user ratings confidence as the reciprocal of entropy value. Then user ratings confidence can be calculated as follows:

$$E_u = \frac{1}{-\sum_i |d_i| \times p(d_i) \log_2 p(d_i)} \quad (1)$$

$$d_i = r_{u,i} - r_i \quad (2)$$

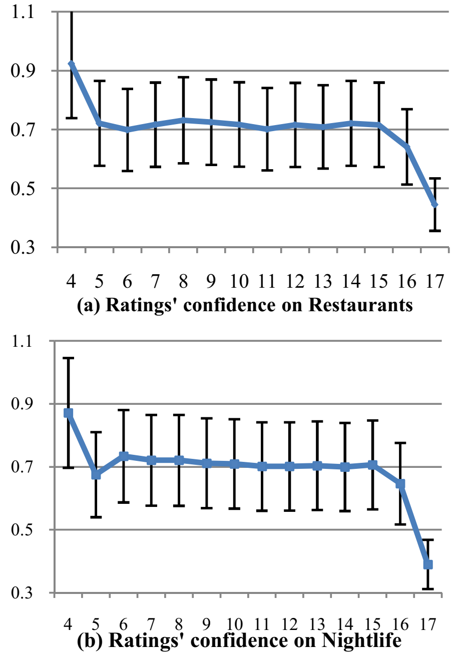
where E_u denotes user u's confidence value, d_i denotes the error value between user rating and item real star level, $r_{u,i}$ denotes user rating and r_i denotes item real star level, $p(d_i)$ denotes the probability of the value of d_i. We can utilize user ratings confidence to evaluate items objectively as follows:

$$\hat{r}_i = \sum_{u=0}^{N} E_u{}^* \times r_{u,i} \quad (3)$$

where $E_u{}^*$ is normalized to unity $\sum_{u \in i} E_u^* = 1$, $u \in i$ means the set of users who has rated item *i*. Note that, *u* is starting from 0. Because there is an additional rating, overall average rating. This idea is to avoid the situation that there is only one rating to test item.

B. Spatial-temporal Features of User Ratings

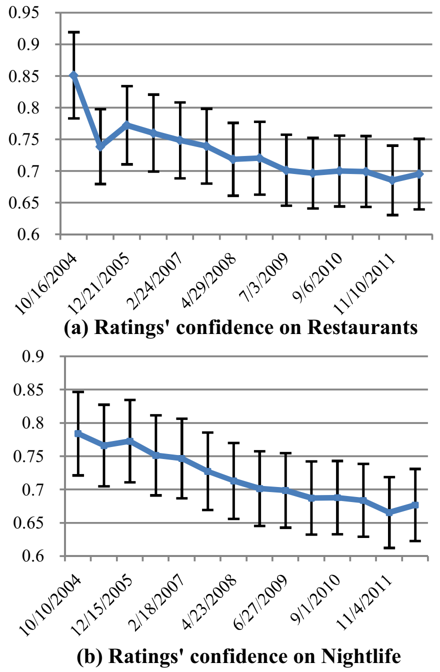
The method of computing user ratings confidence by entropy is based on user overall ratings. That is to say, each user ratings confidence is a constant, whatever the item is. But we know users' profiles are changing constantly. User



(a) Ratings' confidence on Restaurants

(b) Ratings' confidence on Nightlife

Figure 3. The distributions of ratings' confidence in different user-item geographic location distances based on Yelp Restaurants and Yelp Nightlife datasets. The value of x-axis denotes the user-item geographic distance which has been normalized by logarithm, and the value of y-axis denotes the average value of differences between user ratings and item real star levels.



(a) Ratings' confidence on Restaurants

(b) Ratings' confidence on Nightlife

Figure 4. The distributions of ratings' confidence in different periods based on Yelp Restaurants and Yelp Nightlife datasets. The value of x-axis denotes the day time user rated item, and the value of y-axis denotes the average value of differences between user ratings and item real star levels.

ratings confidence may be different in different places at different times. Thus in this part, we focus on each rating's confidence constrained by its spatial and temporal features.

1) Spatial Features

We are living in a large social network, and communicating with diverse people every day. We will be influenced by other people easily, although it may be not our intention. In terms of items, most of them have their competitors. Inevitably there may be some unfair ratings and comments appear on the Internet. Therefore, we suppose that user-item geographic location distance may influence user ratings' confidence.

Fig. 3 is the distribution of ratings' confidence in different user-item geographic location distances based on Yelp Restaurants and Nightlife datasets. The horizontal axis represents user-item geographic distance, which has been operated by logarithm as follows:

$$x = \ln D(u,i) \tag{4}$$

where $D(u,i)$ denotes the distance value between user u and item i. The ordinate axis represents the difference between user ratings and item real star levels, which is an absolute value here. Meanwhile, we also show the proportional standard deviation of each group. We can see that there is not much difference in standard deviations. From Fig. 3, we can see user ratings are mostly unreliable if user is much near to the rated item geographically. As the distance increase, user ratings' confidence is stable. When the distance becomes very large, user ratings are very reliable correspondingly. Why does this happen? We supposes that, users may be influenced by their friends or some discounts of services. In addition, in terms of items, most of them have their competitors. Inevitably there may be some unfair ratings and comments appear on the Internet. Generally speaking, competitors are mostly native. It is reasonable that geographic distance can distinguish different ratings' confidence to a certain extent. We conduct curve fitting to learn ratings' spatial features based on geographic distance. Note that, we conduct curve fitting based on 4th degree Gaussian model according to Fig. 3. In addition we will discuss performance of different fitting curves in section 5. Here its formula is as follows:

$$y = \sum_j a_j \times exp\left(-\left(\left(x - b_j\right)/c_j\right)^2\right) \tag{5}$$

where a_i, b_i, and c_i are the coefficients needed to be learned by curve fitting. We know that ratings' confidence is the contrary of y. Ratings' confidence based on spatial features can be represented as follows:

$$G_{u,i} = 1/\sum_j a_j exp\left(-\left(\left(lnD(u,i) - b_j\right)/c_j\right)^2\right) \tag{6}$$

where $G_{u,i}$ denotes rating's confidence user u to item i. a_i, b_i and c_i are the coefficients learned by curve fitting. $D(u,i)$ denotes the distance value between user u and item i.

2) Temporal Features

In the same way to compute ratings' confidence based on spatial features as described in section 4.2.1, we can get ratings' confidence based on temporal features according to Fig. 4, which shows the distribution of ratings' temporal features in different times. The x-axis represents the rating time, and y-axis represents the difference between user ratings and item real star levels, which is also an absolute value. We show the statistics of temporal features based on our datasets in Fig. 4. Meanwhile, we also show the proportional standard deviation of each group. We can see that there is not much difference in standard deviations. We can see the difference between user ratings and item real star levels is decreasing over time. We suppose that the number of ratings is increasing constantly for each item, which will become a reference to item for other customers. As time passed by, users may get more useful information from former ratings and comments, and give a suitable rating. That is to say, when we search the Internet, we will be unconsciously influenced by the ratings and comments, because the external environment can affect a person's views, especially on the fields he/she does not know well.

Then we conduct curve fitting based on 4th degree Gaussian model according to Fig. 4. We know that ratings' confidence is the contrary of curve. Ratings' temporal features can be represented as follows:

$$T_{u,i} = 1/\sum_j a_j exp\left(-\left(\left(Day(u,i) - b_j\right)/c_j\right)^2\right) \quad (7)$$

where $T_{u,i}$ denotes rating's confidence user u to item i according to temporal features. $Day(u,i)$ denotes the rating time of user u to item i. a_i, b_i and c_i are the coefficients needed to be learned by curve fitting.

C. Proposed Service Objective Evaluation Model

The overview of our proposed service objective evaluation model is shown in Fig. 5. From Fig. 5, we can see the basic idea is that combining user's confidence with spatial-temporal feature to calculate an overall confidence value of a rating. Note that we define confidence coefficient in an effective interval with a constraint that the sum of coefficients is 1. Our present goal is to learn temporal and spatial coefficient vectors of use ratings, because different users' ratings have different coefficient vectors, e. g. a user's rating has high weight coefficient at time t_1 but another one's rating may have low weight coefficient at the same moment. Thus, we aim at learning users' coefficient vectors as the goal of rest work by training them in a unified probabilistic model. Note that, shown in Fig. 5, we set the dimension of statistical chart as the dimension of feature vectors and coefficient vectors.

In order to simplify our formulas, as shown in Fig. 5, we define the overall confidence of the rating user u to item i as follows:

$$\Phi_{u,i} = A_{u,t(u,i)}T_{t(u,i)} + B_{u,g(u,i)}G_{g(u,i)} + C_{u,t(u,i),g(u,i)}E_u \quad (8)$$

Note that:

$$C_{u,t(u,i),g(u,i)} = 1 - A_{u,t(u,i)} - B_{u,g(u,i)} \quad (9)$$

where $t(u,i)$ denotes the time user u rated item i, and $g(u,i)$ denotes the geographic distance between user u and item i. $T_{t(u,i)}$ is the rating's confidence based on temporal features calculated by (7). $G_{g(u,i)}$ is the rating's confidence based on spatial features calculated by (6). E_u is the user ratings confidence calculated by (1). A, B, C are the corresponding coefficient matrixes. These coefficient matrixes sizes are all $M \times k$. M is the number of users, k is the dimension of feature vectors. Then the objective function is given by:

$$\Psi(R, A, B, C, T, G, E)$$
$$= \frac{1}{2}\sum_i\left(R_i - \sum_{u=0}^{n_i}\left(\Phi_{u,i}r_{u,i}/\sum_{u=0}^{n_i}\Phi_{u,i}\right)\right)^2$$
$$+\frac{\lambda_A}{2}\|A\|_F^2 + \frac{\lambda_B}{2}\|B\|_F^2 \quad (10)$$

where the second term is utilized to avoid over-fitting, and $\|\cdot\|_F^2$ denotes the Frobenius norm.

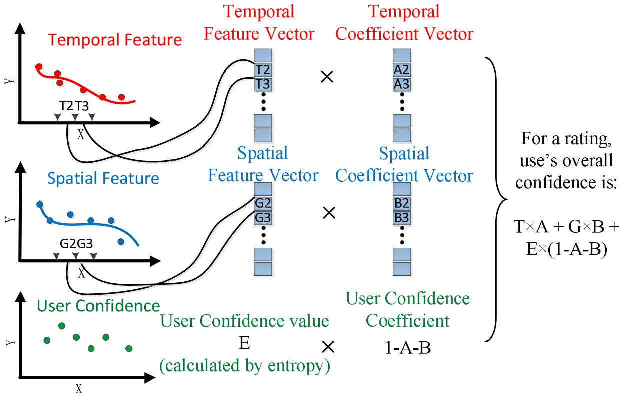


Figure 5. Overview of our proposed Service Objective Evaluation (SOE) model. It deep understands social users by exploring user ratings confidence with considering spatial and temporal features of user ratings.

D. Model Training

Once we get the objective function, we can minimize it by the gradient decent approach as [3], [4], [6], [8], and [9]. The gradients of the objective function with respect to the variables $A_{u,t(u,i)}$ and $B_{u,g(u,i)}$ are respectively showed as (11) and (12):

$$\frac{\partial\Psi}{\partial A_{u,t(u,i)}} = (-1)\left(R_i - \sum_{u'=0}^{n_i}\left(\Phi_{u',i}r_{u',i}/\sum_{u'=0}^{n_i}\Phi_{u',i}\right)\right)$$
$$\times\left(\frac{\left(T_{t(u,i)}-E_u\right)\left(\sum_{u'\neq u}^{n_i-1}\Phi_{u',i}\right)}{\left(\sum_{u'=0}^{n_i}\Phi_{u',i}\right)^2}r_{u,i} + \frac{-\sum_{u'\neq u}^{n_i-1}\Phi_{u',i}\left(T_{t(u',i)}-E_{u'}\right)r_{u',i}}{\left(\sum_{u'=0}^{n_i}\Phi_{u',i}\right)^2}\right)$$
$$+\lambda_A A_{u,t(u,i)} \quad (11)$$

$$\frac{\partial\Psi}{\partial B_{u,g(u,i)}} = (-1)\left(R_i - \sum_{u'=0}^{n_i}\left(\Phi_{u',i}r_{u',i}/\sum_{u'=0}^{n_i}\Phi_{u',i}\right)\right)$$
$$\times\left(\frac{\left(G_{g(u,i)}-E_u\right)\left(\sum_{u'\neq u}^{n_i-1}\Phi_{u',i}\right)}{\left(\sum_{u'=0}^{n_i}\Phi_{u',i}\right)^2}r_{u,i} + \frac{-\sum_{u'\neq u}^{n_i-1}\Phi_{u',i}\left(G_{g(u',i)}-E_{u'}\right)r_{u',i}}{\left(\sum_{u'=0}^{n_i}\Phi_{u',i}\right)^2}\right)$$
$$+\lambda_B B_{u,g(u,i)} \quad (12)$$

Once we get the gradients, we update coefficient matrices as follows:

$$A_{u,t(u,i)} = A_{u,t(u,i)} - \alpha\frac{\partial\Psi}{\partial A_{u,t(u,i)}} \quad (13)$$

$$B_{u,g(u,i)} = B_{u,g(u,i)} - \alpha\frac{\partial\Psi}{\partial B_{u,g(u,i)}} \quad (14)$$

where α is learning rate.

At last, after several iteration computations, we conduct service objective evaluation by learned coefficient matrices as follows:

$$\hat{r}_i = \sum_{u=0}^{n_i}\left(\Phi_{u,i}r_{u,i}/\sum_{u=0}^{n_i}\Phi_{u,i}\right) \quad (15)$$

where $\Phi_{u,i} = A_{u,t(u,i)}T_{t(u,i)} + B_{u,g(u,i)}G_{g(u,i)} + C_{u,t(u,i),g(u,i)}E_u$.

V. Experiments

We implement a series of experiments on Yelp dataset to estimate the performance of proposed methods. Furthermore, we compare the performance of related methods we have mentioned before, including basic method, basic biases, biases based on taxonomy, BaseMF, CircleCon model, PRM, and item-based collaborative filtering. We also discuss the

TABLE III. REAL SERVICE OBJECTIVE EVALUATION EXAMPLES SELECTED RANDOMLY FROM RESTAURANTS DATASET

Given	Item Name	Church's Chicken			Come On In Cafe	Best Burger		Chipotle		Mr Gatti	Boston's Fish House		
	User Ratings	5	1	1	2	2	4	4	5	1	4	5	5
Prediction	BM	2.33			2	3		4.5		1	4.67		
	Biases	2.1			2.01	3.14		4.41		1.01	4.69		
	BT	2.5			2.55	3.36		4.28		2.32	4.10		
	BaseMF	3.77			3.69	3.61		3.89		3.69	3.65		
	CircleCon	3.66			3.46	3.69		3.74		3.65	3.65		
	PRM	3.49			3.40	3.70		3.78		3.64	3.73		
	Item-Based	3.54			3.36	3.92		3.89		3.30	3.83		
	SOE	3.1			3.23	3.40		4.14		2.95	4.05		
Result	Ground Truth	3			3	4		3.5		2.5	4		
	Review Count	36			16	21		87		9	77		

impact of different fitting curves, the impact of data sparsity, the impact of review count, and the impact of each feature.

A. Performance Measures

When we get predicted star levels, the performance of methods will be embodied by the errors. We can see Root Mean Square Error (RMSE) as the most popular accuracy measurements [1]-[7] which are defined as follows:

$$RMSE = \sqrt{\sum_{i\in\mathfrak{R}_{test}}(r_i - \hat{r}_i)^2/|\mathfrak{R}_{test}|} \qquad (16)$$

where r_i is the real star level of item i, $\hat{r}_i$ is the predicted star level. $\mathfrak{R}_{test}$ is the set of all items in the test set. $|\mathfrak{R}_{test}|$ denotes the number of items in the test set.

B. Evaluation

In this section, we compare the performance of our SOE model with other methods, including BM, Biases, BT, BaseMF, CircleCon, PRM, item-based collaborative filtering on Yelp Restaurants and Nightlife datasets respectively.

In Fig. 6, we show the performance based on Yelp Restaurants dataset. We can see that the accuracy of our SOE model is much better than other approaches. Additionally, we find that the performance of matrix factorization models, including BaseMF, CircleCon, and PRM, have little differences on performance. Actually, matrix factorization models are not suitable to solve service objective evaluation, because matrix factorization models aim at personalized ratings prediction. These models focus on computing users' and items' latent feature vectors. In this paper, we firstly utilize them to predict users' personalized ratings, and then average these personalized ratings. It seems inconsistent. Additionally, when we average these personalized ratings, denominator *M*, which denotes the number of users, is so large that the final evaluations have little diversity. That is to say, most of the evaluations we predict by matrix factorization models are in range of 3.4 to 3.9. Thus we can conclude that this usage method of matrix factorization is not suitable to solve service objective evaluation. Moreover, we list some real service objective evaluation examples selected randomly from Restaurants dataset in Table 3. Firstly, we can find that a small amount of ratings cannot represent the overall evaluation of an item. Secondly, we will be confused by the contradictory scores when we refer to former raitngs. It demonstrates the necessity of service objective evaluation.

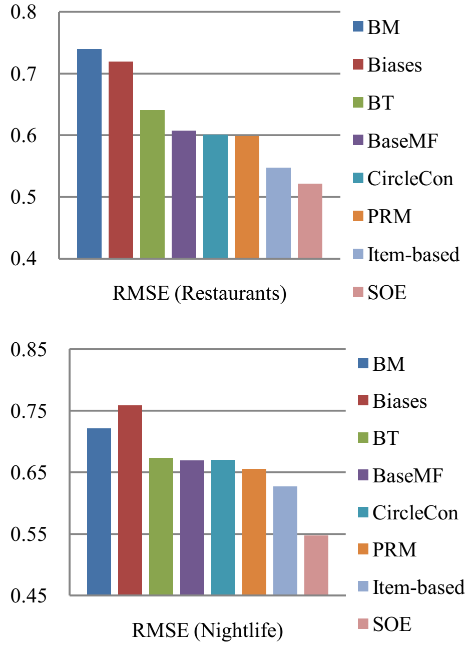


Figure 6. Performance comparison based on Yelp Restaurants, and Yelp Nightlife datasets.

C. Discussion

Besides the performance comparison of the proposed SOE model with the existing BM, Biases, BT, BaseMF, CircleCon model, PRM, and item-based collaborative filtering in Fig. 6, here, we discuss four aspects in our experiments based on Yelp Restaurants dataset: the impact of different fitting curves, the impact of review count, the impact of each feature, and the impact of data sparsity.

1) The Impact of Different Fitting Curves

In this section, we discuss the impact of different fitting curves to performance. As mentioned in section 4.2, we conduct curve fitting based on 4th degree Gaussian model. We conduct series of experiments according to different fitting curves based on Yelp Restaurants dataset as shown in Fig. 7. Note that, P4, P5, P6 denotes fitting curve based on 4th, 5th and 6th degree polynomial respectively. G2, G3, G4

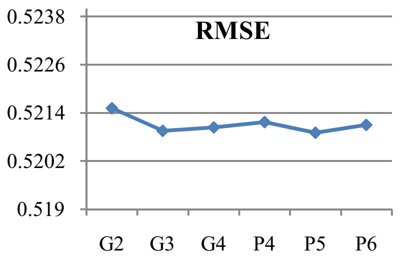


Figure 7. The impact of different fitting curves to performance based on Yelp Restaurants dataset.

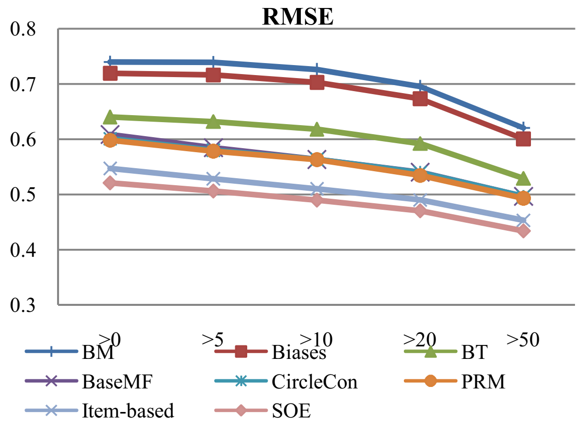


Figure 8. The impact of ground truth to performance based on Restaurants dataset.

denotes fitting curve based on 2nd, 3rd and 4th degree Gaussian model respectively. We find that there is little impact of different fitting curves to the performance. It proves the good robustness of our model.

2) The Impact of Review Count

In this section, we discuss the impact of review count to performance based on Yelp Restaurants dataset. In this paper, our goal is to predict service objective evaluation. But we know that it is difficult to get the ground truth of service objective evaluation, because the ground truth is given by official review site, which is heavily rely on the review count. For example, if the real review count is too small, the ground truths we crawled will be lack of trustworthiness. Thus we discuss the impact of review count by grouping test items. As shown in Fig. 8, we classify testing set into five groups: the real review count of items is greater than 0, 5, 10, 20, and 50 respectively. Intuitively, we deem that performance will become better with the increasing number of real reviewers. This assumption has been proved by our experimental results shown in Fig. 8.

3) The Impact of Each Feature

In this section, we discuss the impact of each feature. As mentioned before, we fuse spatial and temporal features into our SOE model. But we do not know the effectiveness of each feature. Thus, we set user ratings confidence (URC) calculated by entropy in section 4.1 as the baseline. Then we conduct an experiment with considering URC and TF (ratings' temporal features), and another experiment with considering URC and SF (ratings' spatial features). At last, we show the overall performance of our SOE model for comparison. Their performances are shown in Fig. 9, from which we can see the effectiveness of each feature. We can conclude that each feature plays a significant role in proposed model.

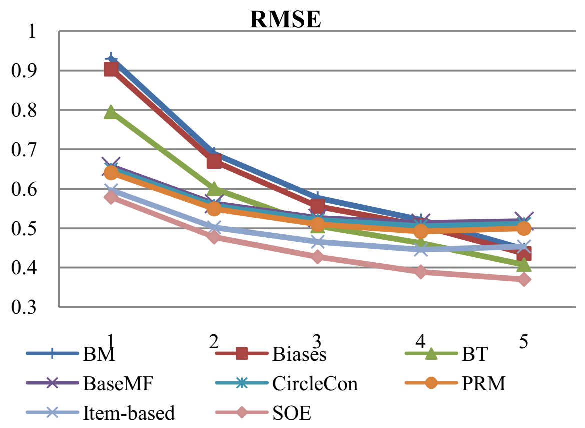


Figure 9. The impact of data sparsity to performance based on Yelp Restaurants dataset.

TABLE IV. THE IMPACT OF EACH FEATURE ON PERFORMANCE

Feature	URC	URT+TF	URC+SF	URC+TF+SF (SOE)
RMSE	0.565	0.538	0.536	0.521

4) The Impact of Data Sparsity

In this section, we discuss the impact of data sparsity to performance. As mentioned before, the number of ratings for each item in test set is no more than 5. Then we conduct series of experiments about impact of data sparsity shown as Fig. 9. According to Fig. 2, we classify Restaurants test set into five groups, with each group just contains the items which has same number of ratings. That is to say, we classify test set into five groups according to different data sparsity. Then from Fig. 9, we find that generally performances are improving with the increase of data density. That is reasonable. But the performances of matrix factorization models are not normal. It has been analysis in section 5.2. We conclude that our model is better than other compared methods in terms of performance, no matter what the data sparsity is.

VI. CONCLUSIONS

Many researchers focus on personalized recommendation and rating prediction. They miss the significance of service objective evaluation, especially for the new services with few ratings. Additionally, local urban services providers can get the feedbacks of their services from world-wide users, which are valuable for them to improve their services qualities. In this paper, we propose an issue of service objective evaluation. To solve the problem of non-objectivity evaluation to items with few ratings, we propose a unified model to evaluate services by deep understanding social users with exploring user ratings confidence. We utilize entropy to evaluate user ratings confidence. Additionally, we find that the spatial-temporal features of users' ratings are helpful to constrain user ratings confidence. Through our model, we can use few ratings to predict star level of item objectivity. Experimental results show outstanding performance of our model. In our future work, we will consider more information for service objective evaluation, such as the sentiments of social users' reviews.

ACKNOWLEDGMENT

This work is supported in part by the Program 973 No. 2012CB316400, by NSFC No.60903121, 61173109, 61332018, and Microsoft Research Asia.

REFERENCES

[1] Y. Koren, "Collaborative filtering with temporal dynamics," KDD'09, pp. 447-456, 2009.

[2] G. Dror, N. Koenigstein, and Y. Koren, "Yahoo! Music recommendations: modeling music ratings with temporal dynamics and item taxonomy," in ACM Recsys, 2011.

[3] M. Jamali, and M. Ester, "A matrix factorization technique with trust propagation for recommendation in social networks," in ACM RecSys, 2010.

[4] X. Yang, H. Steck, and Y. Liu, "Circle-based recommendation in online social networks," in KDD'12, 2012, pp. 1267-1275.

[5] M. Jiang, P. Cui, R. Liu, Q. Yang, F. Wang, W.-W. Zhu, and S.-Q. Yang, "Social contextual recommendation," in CIKM'12, Oct.2012, pp. 45-54.

[6] R. Salakhutdinov, and A. Mnih, "Probabilistic matrix factorization," in NIPS, 2007.

[7] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," Computer, Aug.2009, pp. 30-37.

[8] X. Qian, H. Feng, G. Zhao, and T. Mei, "Personalized recommendation combining user interest and social circle," IEEE Trans. Knowledge and Data Engineering, vol.26, no.7, 2014.

[9] H. Feng, and X. Qian, "Recommendation via user's personality and social contextual," ACM CIKM 2013.

[10] G. Adomavicius, and A. Tuzhilin, "Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions," IEEE Transactions on Knowledge and Data Engineering, pp. 734-749, Jun. 2005.

[11] R. Bell, Y. Koren, and C. Volinsky, "Modeling relationships at multiple scales to improve accuracy of large recommender systems," KDD'07, pp. 95-104, 2007.

[12] B. Sarwar, G. Karypis, J. Konstan, and J. Reidl, "Item-based collaborative filtering recommendation algorithms," In Proceedings of the 10th International Conference on World Wide Web (WWW), pp. 285-295, 2001.

[13] M. Jahrer, A. Toscher, and R. Legenstein, "Combining predictions for accurate recommender systems," KDD'10, pp. 693-702, 2010.

[14] Y. Zhang, B. Cao, and D. Y. Yeung, "Multi-domain collaborative fltering," In Proceedings of the 26th Conference on Uncertainty in Articial Intelligence (UAI), 2010.

[15] G. R. Xue, C. Lin, Q. Yang, W. Xi, H. J. Zeng, Y. Yu, and Z. Chen, "Scalable collaborative filtering using cluster-based smoothing," SIGIR'05, pp. 114-121, 2005.

[16] R. Keshavan, A. Montanari, and S. Oh, "Matrix completion from noisy entries," Journal of Machine Learning Research, 2010.

[17] Y. Koren, "Factorization meets the neighborhood: a multifaceted collaborative filtering model," KDD'08, 2008.

[18] Paterek, "Improving regularized singular value decomposition for collaborative filtering," KDDCup, 2007.

[19] M. Balabanovic and Y. Shoham, "Fab: content-based, collaborative recommendation," Communications of the ACM, pp. 66-72, Mar.1997.

[20] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl, "Evaluating collaborative filtering recommender systems," ACM Transactions on Information Systems (TOIS), pp. 5-53, Jan.2004.

[21] J. Wang, A. P. de Vries, and M. J. T. Reinders, "Unifying user-based and item-based collaborative filtering approaches by similarity fusion," SIGIR'06, 2006.

[22] N. N. Liu, M. Zhao, and Q. Yang, "Probabilistic latent preference analysis for collaborative filtering," CIKM'09, pp. 759-766, 2009.

[23] Q. Liu, E. Chen, H. Xiong, C. Ding, and J. Chen, "Enhancing collaborative filtering by user interest expansion via personalized ranking," IEEE Transactions on Systems, Man, and Cybernetics-Part B (TSMCB), pp. 218-233, Feb.2012.

[24] Y. Chen and J. Canny, "Recommending ephemeral items at web scale," SIGIR'11, pp. 1013-1022, 2011.

[25] M. Harvey, M. J. Carman, I. Ruthven, and F. Crestani, "Bayesian latent variable models for collaborative item rating prediction," CIKM'11, pp. 699-708, 2011.

[26] H. Ma, I. King, and M. R. Lyu, "Learning to recommend with social trust ensemble," SIGIR'09, 2009.

[27] M. Jamali and M. Ester, "Trustwalker: a random walk model for combining trust-based and item-based recommendation," KDD'09, pp. 397–406, 2009.

[28] X. Yang, Y. Guo and Y. Liu, "Bayesian-inference based recommendation in online social networks," The 30th Annual IEEE International Conference on Computer Communications (INFOCOM), 2011.

[29] J. Huang, X. Cheng, J. Guo, H. Shen, and K. Yang, "Social recommendation with interpersonal influence," In Proceedings of the 19th European Conference on Artificial Intelligence (ECAI), pp. 601–606, 2010.

[30] H. Ma, H. Yang, M. R. Lyu, and I. King, "Sorec: Social recommendation using probabilistic matrix factorization," CIKM'08, 2008.

[31] H. Ma, D. Zhou, C. Liu, M. R. Lyu, and I. King, "Recommender systems with social regularization," In ACM International Conference on Web Search and Data Mining (WSDM), 2011.

[32] F. Liu and H. J. Lee, "Use of social network information to enhance collaborative filtering performance," Expert Syst. Appl., pp. 4772-4778, 2010.

[33] P. Massa and P. Avesani, "Trust-aware recommender systems," RecSys'07, pp. 17-24, 2007.

[34] L. Yu, R. Pan, and Z Li, "Adaptive social similarities for recommender systems," RecSys, 2011.

[35] P. Bedi, H. Kaur, and S. Marwaha, "Trust based recommender system for semantic web," IJCAI'07, pp. 2677-2682, 2007.

[36] J. O 'Donovan and B. Smyth, "Trust in recommender systems," IUI'05, pp. 167-174, 2005.

[37] J. Leskovec, A. Singh, and J. Kleinberg, "Patterns of influence in a recommendation network," Lecture Notes in Computer Science, pp.380-389, 2006.

[38] P. Cui, F. Wang, S. Liu, M. Ou, S. Yang, and L. Sun, "Who should share what? item-level social influence prediction for users and posts ranking," SIGIR, pp. 185-194, 2011.

[39] A. Java, X. Song, T. Finin, and B. Tseng, "Why we twitter: understanding microblogging usage and communities," In WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, pages 56–65, NY, USA, 2007. ACM.